A Primer on Probability

Jonathan Shapiro

September 23, 1999

1 What is Probability and Why Do We Use It?

Probability is a notion which we use to deal with uncertainty. If an event can have an number of outcomes, and we don't know for certain which outcome will occur, we can use probability to describe the likelihood of each of the possible events. The classic example is flipping a coin. There are two possible outcomes: the coin could come up heads or tails. Since we don't know which will occur, we say that there is a probability for each of the two events. Probabilities are denoted as numbers between 0 and 1, with 0 meaning the outcome definitely will not occur and 1 meaning it definitely will occur. You expect that if you flip a coin it is equally likely to come up heads or tails; the probability is $1/2$ for each outcome.

However, probability is not just any numeric expression of uncertainty between 0 and 1. It is defined in a particular way with specific rules of calculation. These rules will be discussed below. Probability calculus is not the only formalism for expressing uncertainty, but it is the most important. It is of fundamental importance in scientific reasoning, and is widely used to model uncertainty and random processes in scientific and engineering applications.

1.1 Why Is Probability Important for This Course

In much of this course, we will be concerned with systems which process or classify images or patterns. Why is probability theory an appropriate language for describing such systems? There are several reasons for this. First, in some instances a problem might truly have a random component. An example might be image classification in which some noise somewhere in the image processing system might change the values of some of the features in a random manner. If we want to classify a very noisy image, we may not be able to say with certainty what the original image was; we might be able to express the probability of possible images. Our classification of the image will also be uncertain, so it will be described probabilistically.

More often than not, however, probability is used because we have incomplete information. We use probability to model our ignorance. As an example of this, consider handwritten character recognition. The interpretation of each character might not be random; the writer might always use the same symbol to mean the same thing. However, if we do not know what that is, we have to guess. More formally, we imagine that if we do experiments with the same amount of uncertainty in our information, we would get different results.

An additional source of uncertainty comes with systems which learn from examples. The typical situation is one where learning is done from a tiny fraction of the the possible patterns which the system may see in practice. Thus, the system must learn about the universe from a small sample of examples. This sampling introduces a number of uncertainties which can be expressed probabilistically. For example, it will not be possible to infer precisely the performance of such a system from the number of errors on a small sample, but one can derive the probability of a given error rate. The uncertainties caused by sampling will be quite important in the performance and validation of learning systems.

Learning from examples can be seen as using examples to estimate the probability of classifications. This view is taken in the first part of this course. We will need to express probabilistic ideas mathematically, and require some of the rules for manipulating probabilities. In what follows, I outline basic probability theory.

2 Basic Ideas

2.1 Discrete Probabilities and Probability Distributions

The simplest form of probability functions are discrete probabilities. In this case, we have some variable X which can take different values or outcomes, $x_1, x_2, \ldots$.

Examples: if X represents a statement such as "it will rain tomorrow", it can take two values TRUE and FALSE; if X represents the result of a coin flip, there would be the two possible outcomes, x_1 meaning heads and x_2 meaning tails; if X is the number drawn in the next lottery draw, there would be 13,983,816 possibilities.

We denote the probability that X takes a particular value x_i as $P(x_i)$. The probability must be between 0 and 1. If $P(x_i) = 1$, X takes the value x_i with certainty; if $P(x_i) = 0$, X definitely does not take the value x_i; the closer is $P(x_i)$ to 1, the greater the likelihood that X takes this value. Assuming that X can take only one of the values at a time (the values are mutually exclusive), the probability that X takes x_1 or x_2 is $P(x_1) + P(x_2)$. Since X must take some value, the sum of the the probabilities of all possibly values is 1. Thus, the probability is *normalised*,

$$\sum_{\text{all } x_i} P(x_i) = 1$$

which means that X always takes one of its allowed values.

We are often interested, however, in cases in which a random variable takes a continuous range of values. In this case, the above must be modified. We write a probability *distribution* or *density* $P(x)$ as a function of the continuous variable x.

Example: A farmer might be interested in the probability that a given amount of rainfall will occur in a particular growing season. The rainfall, X, could be modelled by a continuous positive real number x.

We cannot interpret $P(x)$ as the probability that X takes the value x, because that probability is almost always going to be 0. You cannot have an infinite number of possibilities with nonzero probability, because it could not be normalised. For continuous distributions, $P(x)$ is defined so that the probability that X is between x and $x + dx$ is $P(x)dx$, where dx is arbitrarily small. The normalisation condition converts from a sum to an integral, and is,

$$\int_{\text{all } x} dx P(x) = 1.$$

2.2 Combining Probabilities

If there are two variables, X and Y, we might want combined probabilities involving both X and Y. There are three ways of combining random variables.

Joint Probability: First, we can ask for the probability that X takes the value x *and* Y takes the value y. This is denoted $P(x, y)$ (also written $P(x \text{ and } y)$ or $P(x \cap y)$) and is called the joint probability of X and Y. This is simply the probability that both things occur: X is x and Y is y.

Conditional Probability: Another quantity is the conditional probability that X takes the value x *given* that Y takes the value y. This is denoted $P(x|y)$. This is the probability that X takes the value x assuming that Y definitely takes the value y (no matter how unlikely it is that Y takes the value y).

The relationship between the two is

$$\boxed{P(x, y) = P(x|y)P(y)}$$

the probability of x and y is the probability of x given y times the probability of y.

Bayes Rule: From the formula relating joint probability with conditional probability, Bayes Rule follows, since

$$P(x|y)P(y) = P(x, y) = P(y|x)P(x),$$

so,

$$\boxed{P(y|x) = P(x|y)P(y)/P(x).}$$

Note that if X and Y are independent, $P(x, y) = P(x)P(y)$, the joint probability is the product of the independent probabilities.

Example: An example may illustrate the difference between the joint probability and the conditional probability. Suppose we pick an arbitrary British adult. We wonder about two things: is this person rich and did they attend public school? Thus, we have two variables:

- $X =$ the person is rich,
- $Y =$ the person went to public school.

The variables X and Y can each take two values, TRUE and FALSE. Presumably, the probability that the person is rich is small, as is the probability that they attended public school, because there are 10s of millions of British adults, few of them are rich, and few of them attended public school. The joint probability, the probability that the person is both rich and an alumnus of public school is smaller still, since that is the fraction of British adults who are both rich and public school grads.

Now, consider the conditional probability, $P(X = true|Y = true)$. That is the probability that a British adult is rich *given* the fact that they went to public school, i.e. the fraction of public school alumni who are rich. I would assume that this is closer to 1, adults who went to public school are likely to be rich (I assume this is true).

Probability of X or Y: There is a third way of combining X and Y that you should be aware of. That is the probability that X takes the value x *or* Y takes the value y. For this the expression is

$$\boxed{\text{P(x or y) = P(x) + P(y) - P(x,y).}}$$

Thus, if x and y are mutually exclusive, the probabilities simply add.

2.2.1 Conditional Probabilities are of Fundamental Importance

The conditional probability is an essential quantity in wide range of domains, including classification, decision theory, prediction, diagnostics, and other similar situations. That is because one typically makes the classification, decision, prediction, etc. based on some evidence. Thus, what one wants to know is the probability of the result *given* the evidence $P(r|e)$. In the example of classification, the evidence is the values of the measurements, or the features on which the classification is to be based. The possible results are the possible classes. The problem is that this conditional probability is very difficult to estimate from experiments directly. This is because there are typically a huge number of values which the features can take. Thus there would be an enormous number of conditional probabilities to estimate. To get around this, Bayes rule is used to write this in terms of the conditional probability of getting the evidence given the classification. The probability of the evidence conditioned on the result can sometimes be determined from first principles, and is often much easier to estimate. There are often only a handful of possible classes or results. For a given classification, one tries to measure the probability of getting different evidence or patterns. A model is used to interpolate to unseen patterns. This gives an estimate of the conditional probability of the evidence given the classification. Using Bayes rule, we use this to get what is desired, the conditional probability of the classification given the evidence.

3 Rules of Probability Calculation

Here I summarise the rules of probability calculus. They have been stated above; here they are grouped together, so they can be found easily.

1. **Probability is a number between 0 and 1:** $0 \leq P(x) \leq 1$

2. **If the outcome is certain, the probability of that outcome is 1**: If $P(x)$ is the probability for the variable X, $P(x) = 1$ if and only if X takes the value x.

3. **Definition of conditional probability:** The conditional probability is defined by this formula

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

4. **The probability that a variable takes one of a set of mutually exclusive outcomes is the sum of the probability of the outcomes:** $P(x_1 \text{ or } x_2 \text{ or } \ldots \text{ or } x_n) = P(x_1) + P(x_2) + \ldots P(x_n)$ if X can take no more than one of the values $x_1, \ldots, x_n$.

5. **The probability of independent events is computed by multiplying the probability of each event:** $P(x,y) = P(x)P(y)$ if X taking the value x does not depend whether or not Y takes the value y.

6. **Bayes Rule:**

$$P(y|x) = P(x|y)\frac{P(y)}{P(x)}$$

7. **Bayes Rule (alternative form):**

$$P(y|x) = P(x|y)\frac{P(y)}{\sum_i P(x|y_i)P(y_i)}$$

8. **Distribution of a single variable from the joint distribution:** One can find the distribution for X from the joint distribution of X and Y by summing over all values of Y,

$$P(x) = \sum_{\text{all y}} P(x,y).$$

The single distribution determined from the joint is called the marginal distribution; this process is called marginalisation. Note: It is not possible to infer the joint distribution from the marginals.

4 What is the Meaning of Probability?

What is the meaning of probabilistic statements; how do we interpret the statement that the probability of a flipped coin coming up heads is 1/2? There are two interpretations. The most commonly used interpretation is that the probability of an event is the expected frequency that the event occurs. In other words, if you were to do an experiment many times, the average number of times which you get a particular result is the probability of that result (more strictly, this ratio tends to the probability of the result in the limit of a large number of experiments). Probability used in this way is called *objective* probability, as it is based on the view that probability can be measured objectively via experiments. This is also called the frequency interpretation. The objective interpretation is appropriate if there is a definite probabilistic process which generates the outcomes. If you do a number of experiments, the number of times an outcome occurs divided by the number of experiments is an estimate of this true probability of the process. The estimate approaches the true probability as the number of experiments goes to infinity.

For example of coin flipping, you may believe that the process is fundamentally random with fixed probability. To estimate the probability of a coin coming up heads, flip it 100 times, say, and take ratio of the number of heads to the total number of flips. For any particular trial, this would not come out exactly 1/2. Sometimes it would come out 52/100, sometimes 49/100, and so forth. However, as the number of flips got bigger and bigger, this ratio come closer and closer to 1/2, if that is the true probability.

Objective probabilities can be computed as frequencies. A common situation is one in which you would like to compute the probability of a particular outcome x_1. Suppose there are n possible outcomes, all equally likely, and n_1 possible ways of getting the outcome x_1. Then

$$P(x_1) = \frac{n_1}{n}.$$

Example: What is the probability of rolling a pair of dice so that the dots on the upward faces total to 7?

There are 36 different rolls of the dice. There are 5 ways to total to 7. Thus, assuming that each outcome is equally likely, the probability is 5/36.

There is a second interpretation of probability, called *subjective probability.* It will not be used much in this course, but you may see it in later courses on artificial intelligence. This interprets probability as degree of belief. In this interpretation, when I say that the probability that a flipped coin will come up heads it 1/2, I am not describing the results of any experiments (necessarily). I am saying that I believe that the coin is equally likely to come up heads as tails. This interpretation makes sense of situations in which there may be no way of getting an estimate of a "true" probability by repeated trials. As examples, a doctor may say that the probability of a treatment curing you is 80% or a meteorologist may say that the probability of rain tomorrow is 60%. It is not possible to generate many sick copies of you and measure the fraction cured, nor are we able to generate many copies of today and measure the frequency of rainy tomorrows. These experts are telling you something about the strength of their belief in a given outcome.

You may wonder why belief should behave like frequencies. Why should the same rules of calculation be used for belief as for objective probability? There is an argument which says that it is sensible for beliefs to behave this way. Take the case of the doctor. The doctor may require a very accurate measure of the likelihood of the outcome, because decisions will be made on the basis of it. For example, the amount of time and money will the doctor be willing to spend on treating you may depend on this likelihood. Thus, the doctor's beliefs may be measurable from this behaviour. If the doctor's beliefs satisfy the laws of probability, the doctor may spend money and time sensibly; if the beliefs violate basic probability rules, the doctor might not use time and money sensibly. (This notion is usually stated in terms of betting behaviour. The doctor's belief's can be determined by setting betting odds with the doctor; if these beliefs violate rules of probability, you can design a set of bets which will cause the doctor to lose money no matter what the outcomes. This is a so-called Dutch Book. See the text "Artificial Intelligence", by Ian Pratt, for a good exposition of this view.) Thus, probability theory gives a formalism for quantifying beliefs in a sensible way. More than that, it gives a way of combining beliefs with objective evidence.

Viewed this way, Bayes rule can be taken as a way of updating beliefs due to new evidence. If b is a belief in some classification or prediction, etc and x is the evidence, Bayes rule is expressed as

$$P(b|x) = P(x|b)P(b)/P(x).$$

The prior, $P(b)$ is your prior belief, that is, what you think the likelihood of each possibility is prior to receiving the evidence. This formula tells how you should change your belief after receiving the evidence x. What might seem to be nice about this is that you can give a computer or an automated system "beliefs", and those beliefs can change based on new evidence. Whether humans actually update belief in this way is another matter. Whether humans *should* update belief in this way is another matter again.

5 Estimating Probabilities from Data — Statistics

In probability textbooks, you are often given examples in which the underlying probability model is known, and you have to give the likelihood of some data. For example, you might be asked to calculate the probability of flipping 18 heads in a row. It is known (or assumed) that the probability of a coin coming up heads is 1/2, and it is a simple application of the rules for combining probabilities. (Specifically, each flip is independent so the probabilities multiply to give $1/2^{18}$.)

However, often one is faced with the inverse situation — there is some data from which you would like to infer the underlying probability model. This is the problem tackled by "statistics".

Examples: Estimate the probability of getting CJD from eating beef as a distribution function of the amount of beef consumed from a sample of beef eaters; estimate the probability of a computer program performing correctly from test runs.

The problem of the estimation of discrete probabilities is easiest and is dealt with first. The problem of estimation of a probability distribution of continuous variables is important in Bayesian classification, but only the most elementary methods are given here.

5.1 Estimating Discrete Probabilities

Estimating of probability of discrete states is easiest. Given the interpretation of probability discussed in section 4, it is reasonable to use the frequency of the outcome in the sample. In other words, the estimate of the probability of the outcome x_i is

$$\hat{p} = n_i/n,$$

where n_i is the number of times the outcome x_i occur in the sample, n is the total number of points in the sample, and the ˆ over the p is to denote that it is an estimate of the true (unknown) probability p.

Example: Consider flipping a thumbtack (a flat-headed drawing pin, like you would put on someone's chair during your school days). It can land with the point pointing up or with the point touching the ground. What is the probability that it lands pointy side up? To answer this, we can do some experiments. Suppose we flip a thumbtack 20 times, and it comes up pointy side up 13 times and pointy side down times. Our estimate of the probability of a thumbtack landing with pointy side up would be $13/20 = 0.65$.

It is important to emphasise that this is only an estimate of the true probability. Since there will always be fluctuations and variations in a finite sample, it is possible that the true probability is different from this estimate. For example, if the true probability was 1/2 it would be possible to get this data, although this data would be less likely than data in which the frequency is closer to 1/2. Our estimate is the best estimate from the data in the sense that the data would be less likely if the probability was different. Thus, in a sense, this estimate is the mostly likely value given the data. For this reason, it is called a *maximum likelihood* estimate. This can be made more precise, see section 7.

5.1.1 The Uncertainty in the Estimate of the Discrete Probability

There are several was to compute how good an estimate this is. First, we can put an error bar on the value. The standard way of computing the size of the error of an experimental measurement is use what is called the *standard error in the sample mean*, which is the standard deviation of the mean. This is the measured standard deviation divided by the square root of the number of samples. Since this is a binomial process, either the outcome is x_i or it isn't, the formula for the standard deviation of a binomial gives the formula for the standard error, which is,

$$\mathrm{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}.$$

Here N is number of points in the sample. Thus, as the size of the sample increases, the uncertainty of the estimate decreases like one over the square root of the size of the sample. In the example above, our estimate for the probability that a flipped thumbtack lands pointy side up would be 0.65 ± 0.11.

Another approach is to use what are called *confidence intervals*. Given the estimated value of $\hat{p}$, a confidence interval is an range of values in which the true value of p is likely to be. By "likely" one often means that the probability that the true value p falls in the interval is 95%. This is called the 95% confidence interval. You might know that for a normal distribution it is expected that the data falls within one standard error 68% of the time, and within two standard errors about 95% of the time. One says that one has 95% confidence that the true value is between the estimate minus two standard errors and the estimate plus two standard errors. Now, the distribution of the measured estimate is not normal, it is binomial, but a normal distribution can be approximated by a normal distribution if the value of $\hat{p}$ is not too close to 0 or 1. Or one can use a binomial table. Figure 1 shows a graph of the 95% confidence intervals.

In the example about, we would be 95% confident that the true value for the probability of a flipped thumbtack landing pointy side up is between 0.44 and 0.86. If a more accurate estimate is desired, a larger number of experiments is required. The size of the interval will decrease with the square root of the number of experiments.

5.2 Application to Validation and Testing of a Classifier

The most important application of the above from the point of view of this course is the estimation of the error rate of a classifier. This is sometimes called validating or testing the classifier or measuring its generalization performance.

Figure 1: Confidence intervals for a binomial variable. $\hat{p}$ is the measured value; p is the true value. The curves are labelled by the number of sample points.

Suppose we have a classifier which has been developed using learning techniques or any other method. We can view the error-rate as the probability that the classifier will misclassify a pattern. To estimate this, the classifier is applied to a set of test patterns; the fraction of the samples misclassified is used as the estimate of the error rate,

$$\hat{p} = \frac{k}{n},$$

where k is the number misclassified in the sample, and n is the number of patterns in the sample. If data was used to develop the classifier, it is most important that a different set of patterns be used to test it.

We can see from the figure that if n is not large the true error rate can be fairly different from the measured one. For example, if the test set contains 50 patterns and the classifier correctly classifies all of them, at the 95% confidence level the true error could lie between 0 and eight percent. Such reasoning can help you determine how much test data you will need. For example, suppose you are required to produce a classifier which is correct 97% of the time. You will require sample size of about 250, so that if you can get a classifier to misclassify less than one percent of the sample, the true error is likely to be less than three percent. (How to make such a classifier is another problem.)

5.3 Estimating Continuous Probabilities

This is a much more difficult problem than the problem of discrete probability estimation. This is because, if $P(x)$ is a probability distribution on a continuous variable, x, we have to estimate its value at an infinite number of points x. Of course we will only have a finite sample of data points with which to do this. Clearly, some type of interpolation between these points will be required. There is a substantial literature on this problem, which is known as probability density estimation. See, for example, Chapter 2 of *Neural Networks for Pattern Recognition* by Bishop, or chapters 3 and 4 of *Pattern Classification and Scene Analysis* by Duda and Hart for a treatment.

This problem will be important to us because we will often have a continuously varying measurement which will be used to make a classification. Thus, we will be attempting to estimate $P(c|m)$, where c is the classification and m is the value of the measurement. That can be done by estimating $P(m|c)$, that is the probability distribution for the values of the measurements for each classification, which is related to $P(c|m)$ through Bayes rule. The estimation of $P(m|c)$ is one of density estimation.

Methods of density estimation fall into two classes: parametric and non-parametric. Parametric approaches assume a particular form for the probability distribution. There are some parameters which are adjusted to make the distribution most like the data. For example, a common approach for one-dimensional data is to assume that the data comes from a normal or Gaussian distribution. This distribution is specified with two parameters, the mean and the standard deviation. It turns out that the best way to set those parameters is use the mean and the standard deviation of the data. It can be shown that these are the maximum likelihood estimates. The problem with this approach is that the assumed form of the distribution might not be appropriate for the problem.

An example of a non-parametric method is a histogram. To make a histogram, you simply divide the region which contains the data into a number of equal sized sections or *bins*; the estimate of the density for each section is the fraction of the data which falls inside that bin. In other words, to estimate of $P(x)$ for any x, find the bin that x is in, say that is the ith bin, and use the estimate $P \sim N_i/N$, where N is the number of data points used to make the histogram, and N_i is the number of data points in the ith bin.

The choice of the bin size determines how smooth the data appears. If the bin-size is very small, the density will appear very spikey; if the bin size is large, important features of the data might be washed out. One of the problems with histograms is that one never knows how to set the bin size. Another problem is that they can only be used on low dimensional data.

As an example of density estimation, consider the following set of numbers:

37	28	36	42	26	27	37	48	26	19	48	42
33	38	20	11	28	40	48	28	41	44	27	46
17	28	14	44	21	48	29	35	32	14	20	43
18	25	38	24	27	43	46	24				

Suppose we want to model a probability distribution which may have produced the data. First, model the data as a normal distribution. The mean of the data is $m = 32.05$. The standard deviation of the data is $\sigma = 10.7$. Thus, the distribution could be modelled as a normal distribution with these values for the mean and standard deviation. This is shown as the first graph in figure 2 (the data is also shown as diamonds on the bottom of the graph).

To produce a histogram, divide the region which contains the data in to a number of bins. In this case the data lies in the interval 11 to 48. The remaining graphs in figure 2 show histograms for different bin sizes. Notice that the histograms seem to show that the distribution has two peaks. The Gaussian approximation, misses this, of course.

6 Likelihood of Data from Known or Assumed Probability Distributions — Hypothesis Testing

Hypothesis testing is not an important part of our course. However, it places a big part of probability and statistics, so I include a brief mention of it here.

In the above sections, the problem considered was to estimate the probability from some data. A related problem is the testing of hypotheses — that is testing whether some data is consistent with the believed or proposed distribution. Of course, in many cases all data is consistent with the hypothesis, but may be extremely unlikely. The classical approach to this problem was developed by Fisher (1935). First pick a *statistic* which is a measurement of the data which can determine how likely the data is. Second pick a threshold percentage, usually taken to be 95%, and accept the hypothesis if the statistic on that data is within the range which would occur that percentage of the time.

For example, suppose someone flips a coin 20 times and it comes up heads 4 times, tails 16 times. Is this a funny coin? You might think so, but how would you test this? You expect 10 heads, but you

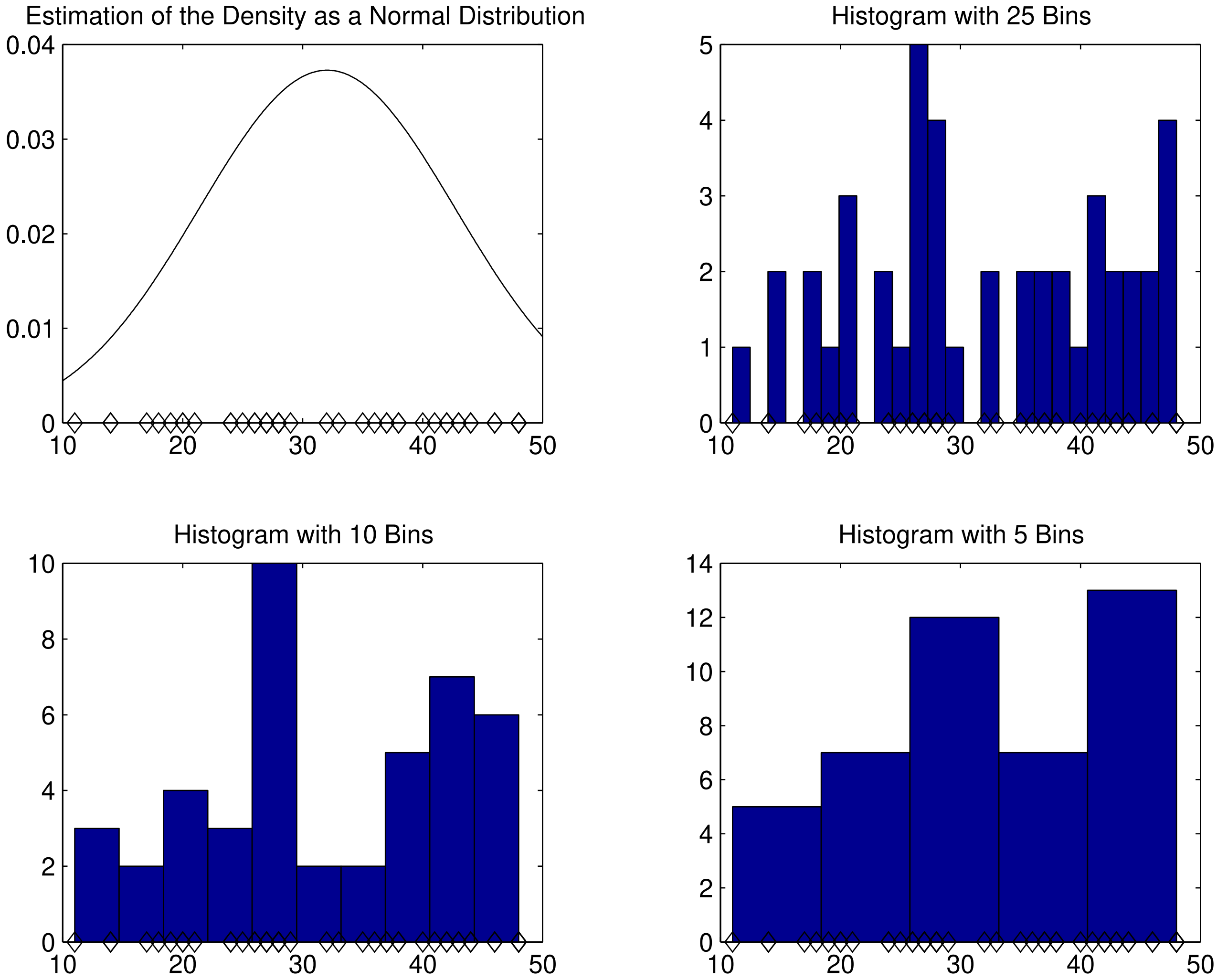

Figure 2: Examples of density estimation. The data is shown as diamonds at the bottom of the graphs. The first graph shows a Gaussian approximation to the distribution. The others show histograms with varying bin sizes.

know this will fluctuate. You take as your hypothesis: the coin is a fair coin; the probability of heads is 50%. This is the so-called *null hypothesis*. Then you ask, is the deviation from the expected number of heads greater than the deviation which you would expect to find 95% of the time? In essence, you considered the confidence interval around the *assumed* value. If the measured value falls within that, the null hypothesis is accepted; if the measured value falls outside this confidence interval, the null hypothesis is rejected. The distribution is binomial, so you can look this up on the curves above. Or you could compute it yourself from properties of binomials. What you would find is that the probability of getting expected value, 10 heads in 20 trials is 18%. The probability of getting within 1 of the expected value, that is 9 heads, 10, heads or 11 heads, is 50%. Table below shows the the percentage of the probability of which is within Δ of the expected value either way for 20 trials of a binomial process with $p = 1/2$.

Δ	allowed # of heads	percent in this interval
0	10	18%
1	9,10,11	50%
2	8,9,10,11,12	74%
3	7,8,9,10,11,12,13	88%
4	6,7,8,9,10,11,12,13,14	96%
5	5,6,7,8,9,10,11,12,13,14,15	98.8%
6	4,5,6,7,8,9,10,11,12,13,14,15,16,	99.7%
7	3,4,5,6,7,8,9,10,11,12,13,14,15,16,17	99.96%
8	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	99.996%
9	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19	99.9998%
10	0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,20	100%

We see that we would not accept the coin as a fair coin at the 95% level.

Classical hypothesis testing is very important in science, and the subject fills books. The procedure is as above, but with several complications. There are different types of hypotheses to test – one-sided versus two-sided, comparisons of two distributions, and others. There are different distributions which used in as the hypothesised process. There are special statistics which come from known distributions in the limit of large data sets for very general processes. It is also worth noting that classical hypothesis testing is problematic in that a given hypothesis can be accepted or rejected with the same data by different procedures or statistics.

7 Inferring Probability Distribution from Data — The Bayesian Approach

In section 5, it was shown how to compute a single estimate for a discrete probability. This estimate could differ from the true value, it is simple the most likely value. Thus, we only know the value of p probabilistically. Bayesian methods provide a way of getting the entire probability distribution for p, not just the most likely value. This approach is explained here. This is an advanced topic, and is an area of considerable current research.

Thomas Bayes (1763) asked the inverse question from that discussed in the previous section. In the above we asked, given a probability distribution, what is the likelihood of a set of data. Bayes asked the question, given a set of data, what can be said about the probability distribution from which it came. Formally, if D represents the data and H represents some hypothesis about a probabilistic process which generated this data, the question being asked is what is the probability of the hypothesis given the data $P(H|D)$? Bayes rule can be used to convert this to the problem above

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

The conditional probability $P(D|H)$ is the probability of a set of data given a know process H which is often straight-forward to calculate (and was discussed in the previous section). The quantity $P(H)$ is called the *prior* – it is the probability of the hypothesis independent of the data. The probability of the data $P(D)$ can be thought of as normalisation and calculated from

$$P(D) = \sum_H P(D|H)P(H).$$

Example: You may have heard the pessimistic adage "dropped toast always lands buttered side down". Suppose you test this. You drop 10 pieces of toast, under controlled conditions of course, and find that it lands butter side down 3 times out of the ten. What is the probability that a dropped piece of toast lands butter side down?

Clearly, this is a binomial (Bernoulli) process. The probability of the toast landing butter side down is p, but you don't know p. Bayes formula can be used to calculate the probability distribution or p given this data. A question is, what should we use as the prior? One possibility is to use our prior belief. Perhaps we believe in the absence of data that the expected value should be 1/2 in which case we could use some distribution which has mean 1/2 and variance which is a measure of our confidence in this prior belief. Another possibility is assume that all ps are equally likely,

$$P(p) = 1.$$

We will use this here (there are arguments against use of a uniform prior, but we shall ignore them as technicalities). Since it is a binomial process,

$$P(d|p) = \binom{10}{3} p^3(1-p)^7.$$

Using Bayes Rule with a uniform prior and normalising, we get a probability distribution for p

$$P(p|d) = 11\binom{10}{3} p^3(1-p)^7.$$

In general, if we drop the toast N times and it lands butter side down n time, Bayes' rule with the uniform prior gives

$$P(p|d) = (N+1)\binom{N}{n} p^n(1-p)^{N-n}.$$

The figure below shows $P(p|d)$ for $N = 10, 20, 40, 80, 160, 320,$ and 640 and $n = 0.3N$ (of course in a real experiment, the measured mean would fluctuate even dropping toast was truly binomial with $p = 0.3$). For all values of N the distribution is peaked around $p = 0.3$, of course. As the number of trials increases, the distribution becomes more and more sharply peaked around this value.

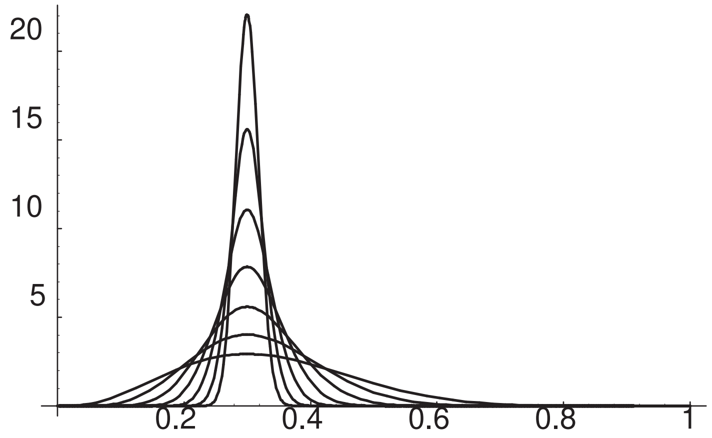

Figure 3: The probability distribution for p for different sample sizes. As the sample size increases, the distribution becomes increasingly sharp. The maximum likelihood estimate is at the peak of the distribution.

Can we conclude from this data that toast always lands butter side down? From this data, you could infer that $p = 0.3$. This is the most likely value; choosing it would a *maximum likelihood* *method*. The Bayesian approach gives you a probability distribution for p. Thus, we could ask, how likely is it that $p \geq \frac{1}{2}$, that is, how likely is it that dropped toast is more likely to land butter side down. This likelihood could be computed from

$$P(p \geq 0.5) = \int_{0.5}^{1} P(p|d)dp.$$

For 10 trials, 3 landing butter side down, this is about 11%. Thus, we would not reject the hypothesis "Dropped toast is more likely to land butter side down" at the 95 % level. However, if there were 20 trials of which 6 landed butter side down, $P(p \geq 0.5)$ is about 4 %; with that data we could reject this hypothesis at the 95 % level.

8 Appendices — Important Probability Distributions

A. The Binomial Distribution

This you have seen in CS1021. It is also called a Bernoulli distribution. It is applicable to the situation where there are N independent trials, each of which can have one of two outcomes. Call one of the outcomes 0 and the other 1. Let p be the probability that any trial yields the outcome 1. Then the probability of getting n 1's in N trials is given by

$$P(n|p, N) = \binom{N}{n} p^n (1-p)^{N-n},$$

where

$$\binom{N}{n} \equiv \frac{N!}{n!(N-n)!}.$$

The expected mean number of 1's is Np; the variance is $Np(1-p)$

In the limit of large N with p fixed, a binomial distribution is well approximated by a continuous Gaussian distribution, with the appropriate mean and variance.

B. The Gaussian or Normal Distribution

There is a very common and important example of a continuous probability distribution. This is the Gaussian, or Normal distribution, also called a bell-shaped curve. It looks like this figure 4.

It is determined by two numbers, the location of the peak and the width. Mathematically, the two parameters which define it are the mean m and the standard deviation σ^2. The mathematical form of the distribution is

$$P(x|m, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$$

. The normal distribution is important because the sum of a set of independent variables drawn from almost any distribution approaches Gaussian as the number of variables goes to infinity. So, for examples, means tend to be Normally distributed. This result is called the central limit theorem.

There is a multidimensional version of the Normal distribution. It is called a multi-variant Gaussian; it is shown in two-dimensions in figure 5.

In this case it is defined by the centre of the peak, two directions of spread and two widths along those two special directions.

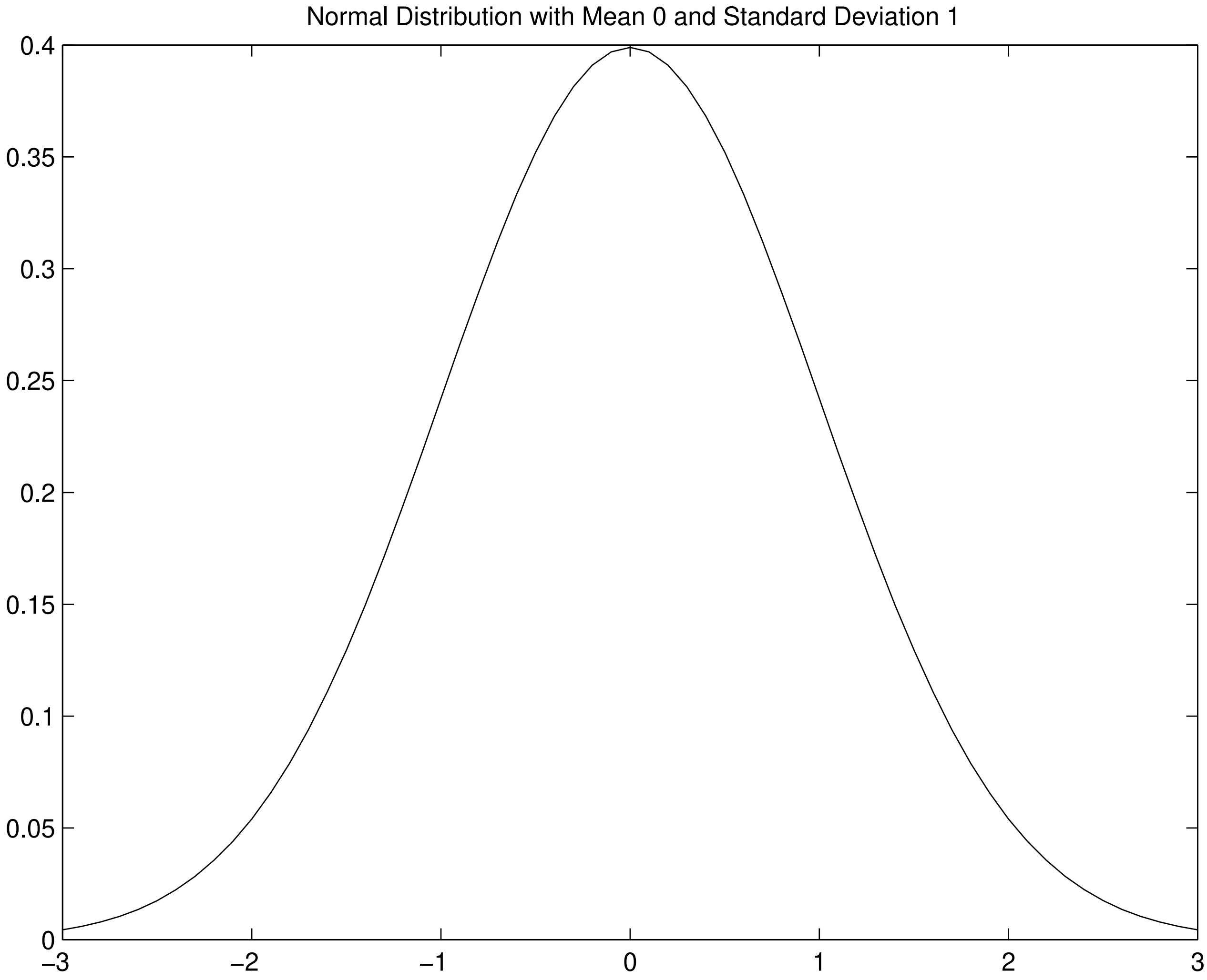

Figure 4: A normal or Gaussian distribution with zero mean and unit standard deviation.

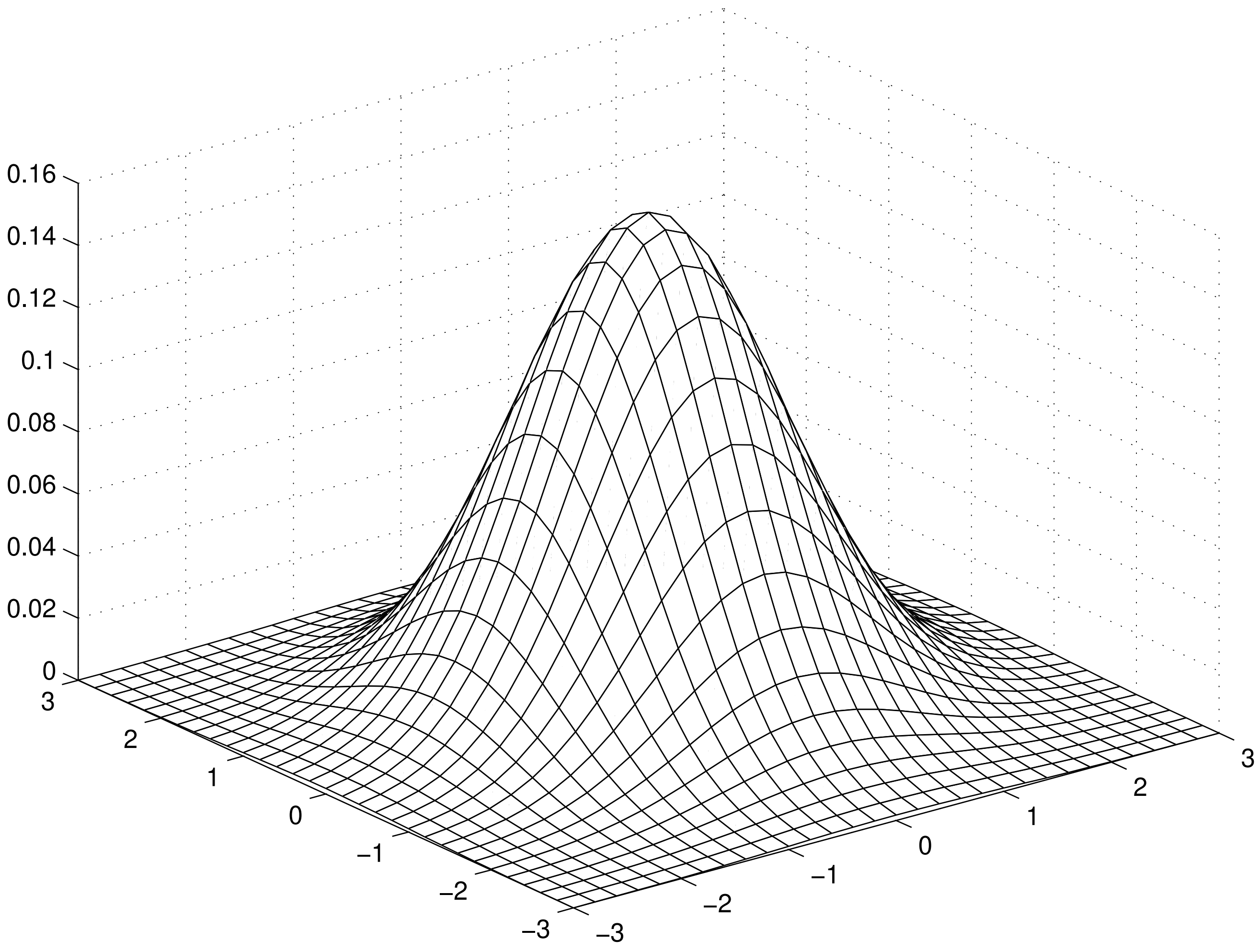

Figure 5: A two-dimensional Gaussian